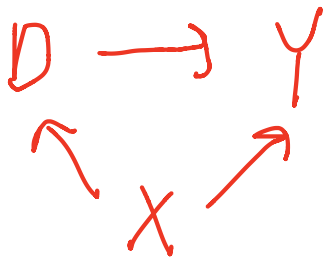
Directed Cyclical Graphs

Introduction

- Graphical models aimed at helping with identifying causal effects in observational data.
- Causal graphs provide a interactive and nonparametric representation of causal relationships. They are a visual representation of the inference problem of interest.
- In Directed Acyclical Graphs (DAG) notation causality generally runs in **one direction**. Causal relationships are represented by directed arrows (single headed arrow).
- DAG uses nodes and arrows to illustrate causal effects. Arrows represent causal effects between random variables.
- DAGs help illustrate the following biases 1) omitted variable bias, 2) reverse causality, 3) incorrect conditioning
- DAGs are motivated from economic theory, assumptions need to be made on the direction of the causal effect.

DAG Ex.



D = treatment
X = Confounder
Y = outcome

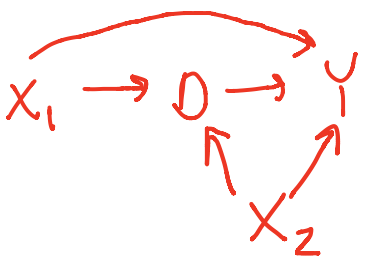
- The DAG above illustrates the standard endogeneity problem when determining the effect of D on Y.
- We are interested in the direct effect of D on Y, but the confounder X can drive spurious relationships between D and Y. In DAG terminology the confounder results in a **backdoor path**, that is a alternative path from D to Y.

Direct: $D \rightarrow Y$, Backdoor: $D \leftarrow X \rightarrow Y$

- To obtain the direct causal effect of D on Y, we can condition on the confounder X. If we don't condition on X, then the backdoor path is **open** and so the direct effect is not identified.
- Conditioning on X **blocks** the backdoor path D <-- X --> Y allowing us to isolate the direct effect D -- > Y.

Condition on $X \Rightarrow D \leftarrow \boxed{X} \rightarrow Y$ is closed

- Notice that box around a random variable in a DAG denotes that it is being conditioned on.
- Let us consider a slightly more complex DAG:



$\Rightarrow$ open Backdoor paths

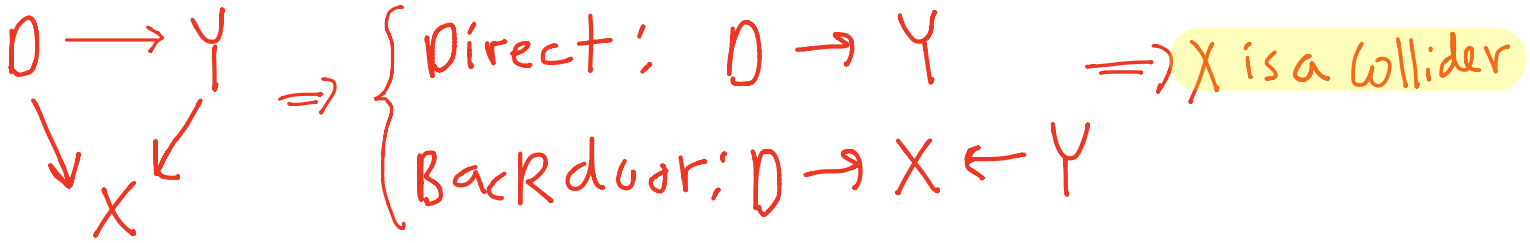
$D \leftarrow X_1 \rightarrow Y$

$D \leftarrow X_2 \rightarrow Y$

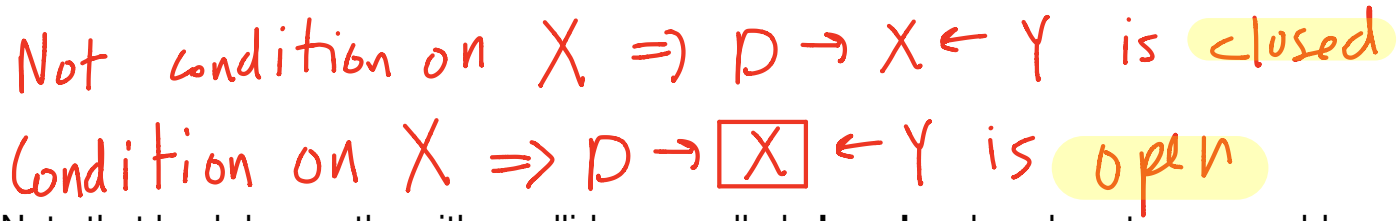
- We are are interested in the direct effect of the treatment on outcome, but identifying this is made difficult by the two **open** backdoor paths listed above.
- Controlling for both X1 and X2 **blocks** both the backdoor paths and allows us to identify D --> Y.

Colliders

- A **collider** is a variable that is caused by both the treatment and outcome.



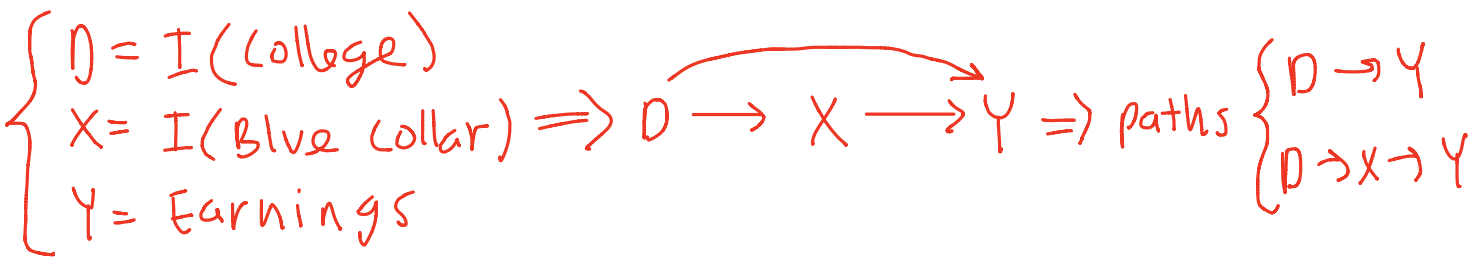
- As long as we don't condition on X, in the above DAG the direct effect D --> Y is identified. This is because the backdoor path is currently **closed** .
- However if we condition on X, that will **open** the backdoor path (initially closed before conditioning) and hence the direct effect will not be identified.



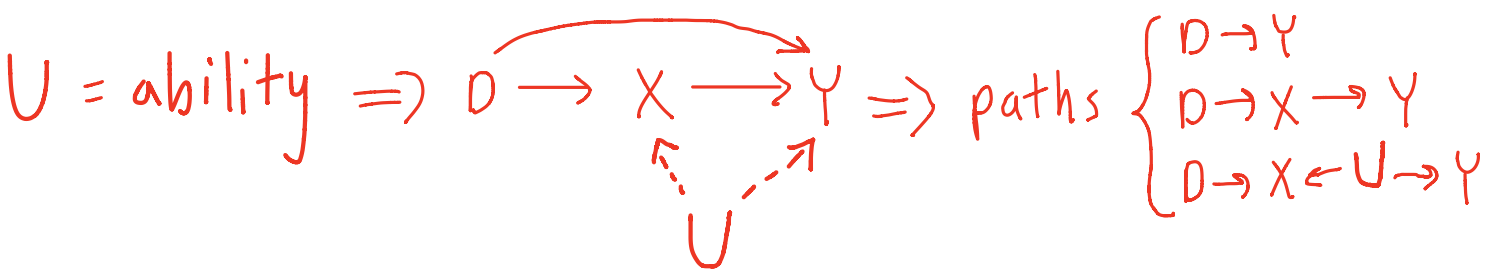
- Note that backdoor paths with a collider are called **closed** and so do not cause problems for identifying the direct effect.
- Similarly backdoor paths with no colliders are known as **open paths**. Having **open** back door paths are problematic for identification of direct effects.
- **Colliders block** indirect relationships (alternative paths from D to Y) between treatment and outcome.
- Therefore to identify the direct effect, we want to close all backdoor paths by conditioning on **non-colliders** (close backdoor paths that are initially open) and not conditioning on colliders (prevent opening backdoor paths that are initially closed).

"Bad Control" vs. Collider Discussion

- **"Bad controls"** (mostly harmless econometrics) are variables that are directly effected by the treatment. Conditioning on a bad control can lead to biases. Let us illustrate this using a DAG:



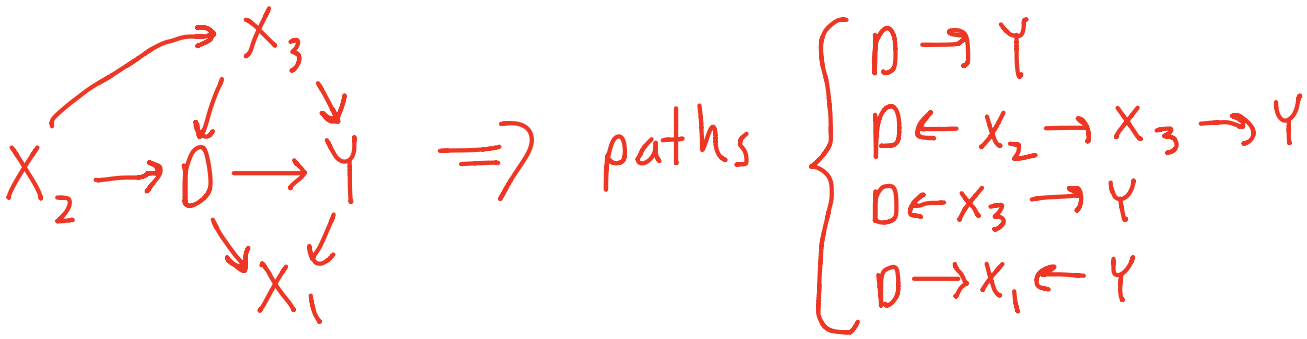
- Conditioning on X in the above DAG fixes the value of the treatment and eliminates the direct relationship between D and Y. That is conditioning on X explains some of the effect of D on Y. Hence the ATE of D on Y is not identified.
- Suppose unobserved ability causes occupation type and earnings:



- Notice the dashed lines denote that U is unobserved, so we cannot control for U.
- X is a **bad control** in the path D --> X --> Y, but is a **collider** for the path D --> X <-- U --> Y.
- Hence if condition on X, then the last backdoor path becomes **open**. We cannot condition on U since it is not observed and this introduces additional biases via U in estimating the direct effect.
- Conditioning on a **bad control** (descendant of the treatment) or a **"collider"** is not recommended as it can lead to a spurious correlation between D and Y.

Backdoor Criterion

- The above discussion on identifying the causal effect of the treatment on outcome (by conditional independence assumption) can be summarized by the **backdoor criterion.**
- The backdoor criterion says that conditioning on a set of variables X identifies the direct path D --> Y if X properly blocks all **open** backdoor paths. This means 1) X doesn't contain **colliders** and 2) X doesn't contain **bad controls**.
- The backdoor criterion essentially gives conditions in which the regression approach can be used to identify causal effects (conditional independence assumption).
- Let us use the backdoor criterion to identify the direct effect in the following DAG:

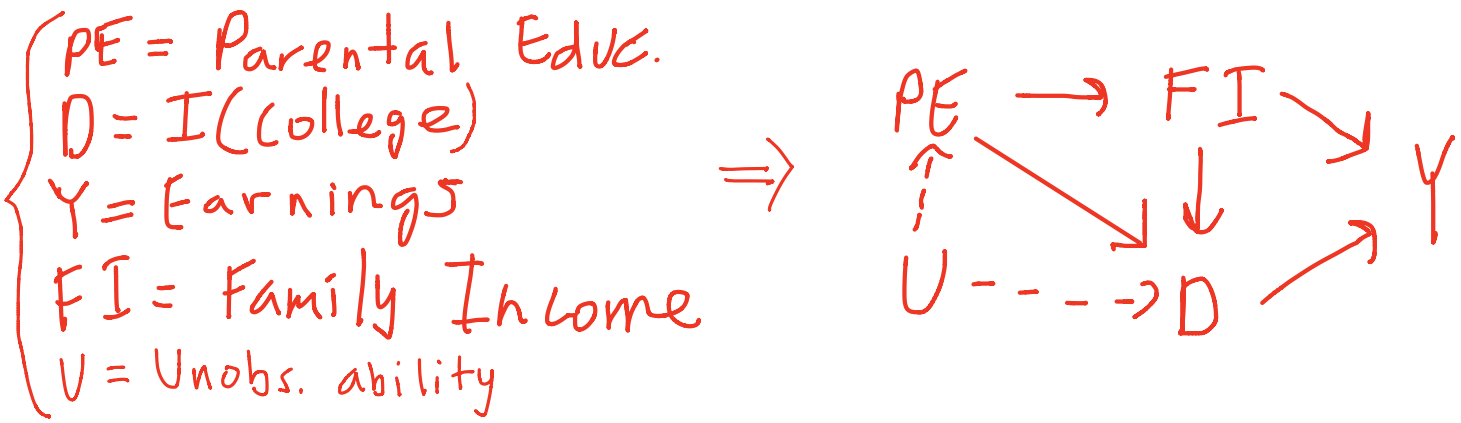


- There are three backdoor paths associated with the above DAG, however the last one is already closed since X1 is a collider along the path D --> X1 <-- Y.
- There are two open backdoor paths. We can close both of them by controlling for X3.
- Hence X3 in this case satisfies the backdoor criterion.
- Note that there are three ways to block the first backdoor path D <-- X2 --> X3 --> Y, we can control for X2, or control for X3, or control for both X2, and X3.

Practical Examples

<u>Returns to Education:</u>

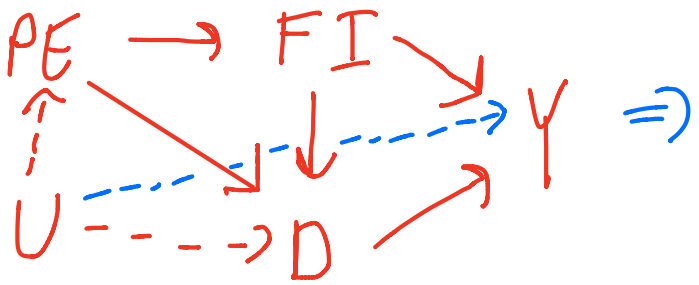
- Let us discuss the popular returns to college education example. Consider the following DAG:



- Note that above DAG assumes individual ability does not directly impact earnings. The graph also illustrates several other assumptions about the relationship of the variables of interest.
- We are interested in identifying the causal effect of going to college on earnings, but there are several backdoor paths that are illustrated below.

Direct: $D \to Y$, Backdoors $\begin{cases} D \leftarrow FI \to Y \\ D \leftarrow PE \to FI \to Y \\ D \leftarrow U \to PE \to FI \to Y \end{cases}$

- In the returns to education example above, controlling for FI (Family Income) satisfies the **backdoor criterion** as FI is a **non-collider** (and also not a bad control) in all the backdoor paths.
- Suppose now we assume U (unobserved ability) had a direct effect on Y:

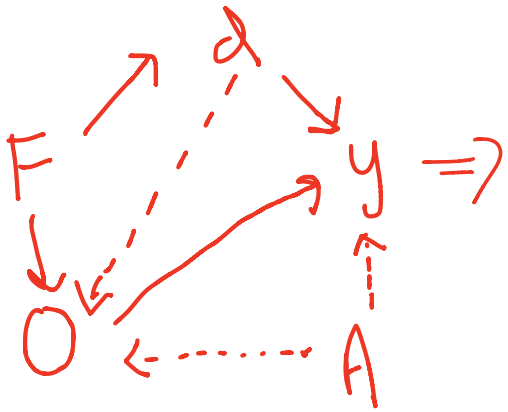


$\Rightarrow$ open backdoor: $D \leftarrow U \to Y$

- We now get another backdoor path of D <- U -> Y. Since U is unobserved, we cannot close this backdoor path.
- If unobserved ability directly effects earnings, then the backdoor criterion cannot be satisfied by just controlling for family income. The returns to education is not identified in this case.

Gender Earnings Bias:

- Let F = I(female), d = discrimination, A = unobserved ability, y = earnings, and O = occupation.
- The following DAG represents the problem of interest:



$\Rightarrow$ open Backdoor $\begin{cases} d \to O \to y \\ d \leftarrow F \to O \to y \end{cases}$

closed Backdoor $\begin{cases} d \to O \leftarrow A \to y \\ d \leftarrow F \to O \leftarrow A \to y \end{cases}$

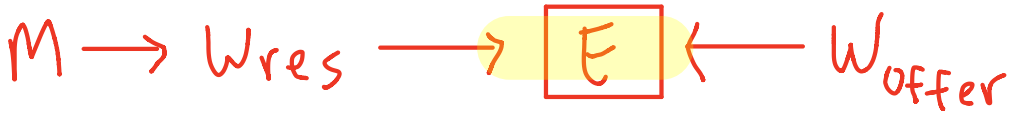
- We are interested in identifying the effect of discrimination (d) on earnings (y).
- Controlling for occupation (O) will close the first two backdoor paths (that are initially open), but will **open** both closed backdoor paths since occupation is **collider** along those paths.
- Even after conditioning on occupation, the last backdoor path (opened after conditioning on just occupation) can be closed by controlling for the female indicator.

Condition using $\{O, F\} \Rightarrow$ $\begin{cases} d \to [O] \to y \\ d \leftarrow [F] \to [O] \to y \\ d \leftarrow [F] \to [O] \leftarrow A \to y \end{cases}$ closed

- However the second last backdoor path d--> O <-- A --> y (also opened after conditioning on occupation) cannot be closed as ability is unobserved.
- Note that occupation (O) is a **bad control** along d --> O --> y, hence conditioning on O is problematic even if we observed ability A.

Sample Selection Problem:

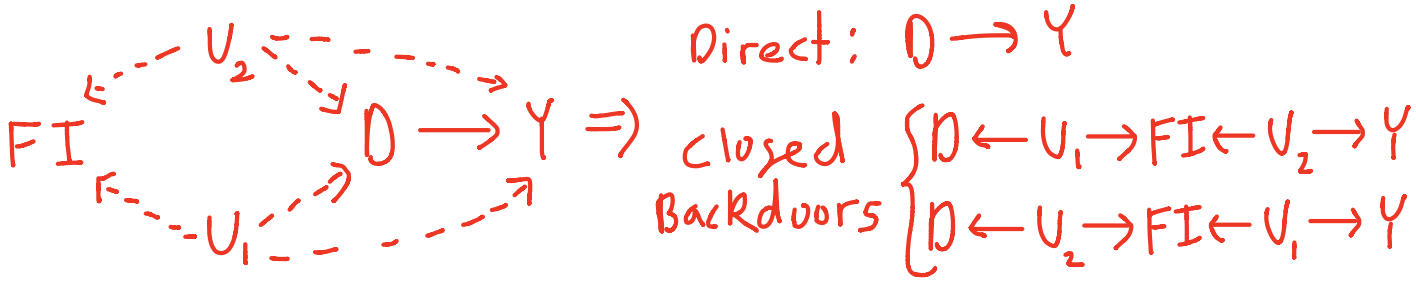
- Sample selection occurs when the participants in the data are not representative of the population of interest.
- Suppose we want to determine the causal effect of motherhood on wages. The sample selection problem is that wages are only observed for employed females.
- Let M = motherhood status, Wres = reservation wage, Woffer = wage offered for job, and E = employment status.
- Let the following DAG represent the problem of interest:



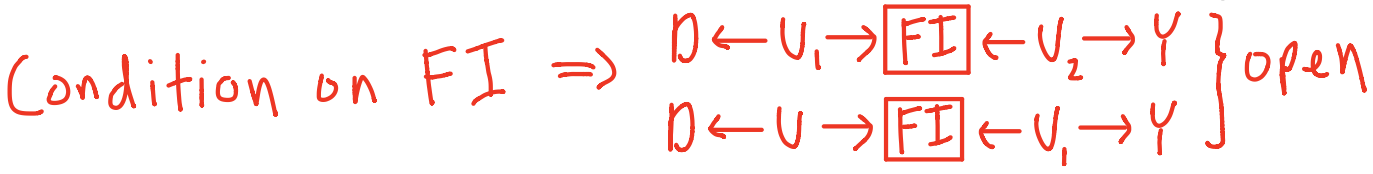
- The assumption made here is that motherhood does not have any effect on wages. However we assume motherhood effects your reservation wage.
- Reservation wage and offered wage are both clearly related to employment (only work if offer is more than the reservation wage).
- The above DAG shows that we are forced to condition of employment (due to sample selection), this is a **collider** in the path M --> Wres --> E <-- Woffer.
- Hence even though motherhood has no effects on wages, the data will falsely indicate that motherhood is related to wages. This is another example of **collider bias**.

Controlling for pre-treatment Characteristics

- Controlling for pre-treatment characteristics is normally considered as a good practice. However it is possible for this variable selection strategy to cause biases.
- Suppose we want to determine the returns to college education.
- Let D = I(college graduate), Y = earnings, U1 = snobs. mother ability, U2 = snobs. father ability, and FI = family income. Now consider the following DAG:



- Since U1 and U2 are unobserved, the paths D <-- U1 --> Y and D <-- U2 --> Y are always **open**. Therefore the returns to education is not identified.
- If we condition on the **pre-treatment** family income, the two backdoor paths **open** and introduce additional biases. This is because FI is a collider in both the backdoor paths.

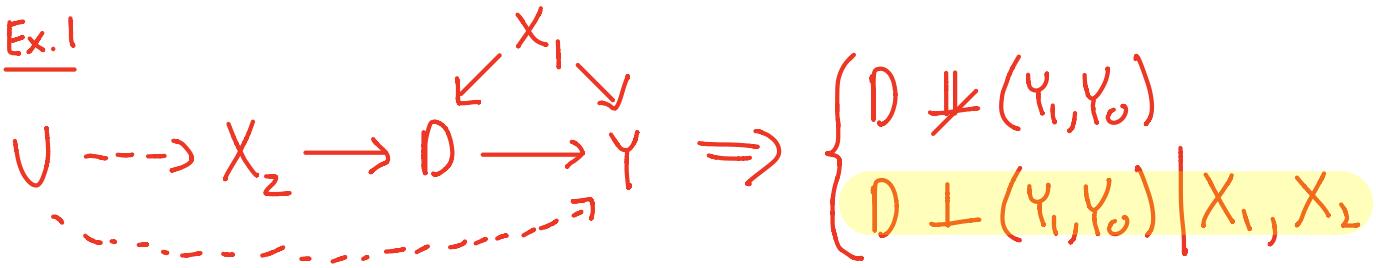


- Key takeaway is that it is still possible for pre-treatment controls to lead to **collider bias.** This type of situation may be rare, here we assumed FI is unrelated to college graduation.

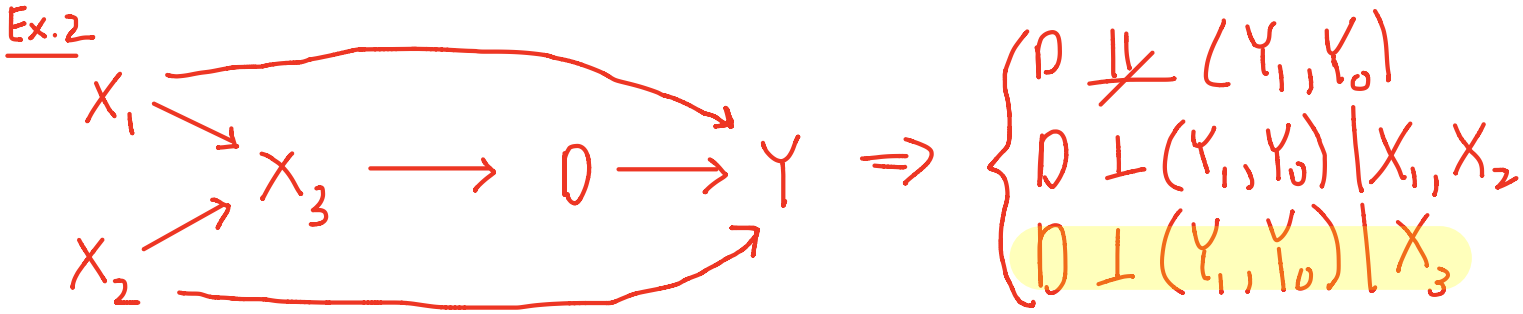
Relationship between DAG and Standard Identification Assumptions

- The potential outcome model says that the average treatment effect is identified if the treatment is independent of the potential outcomes:

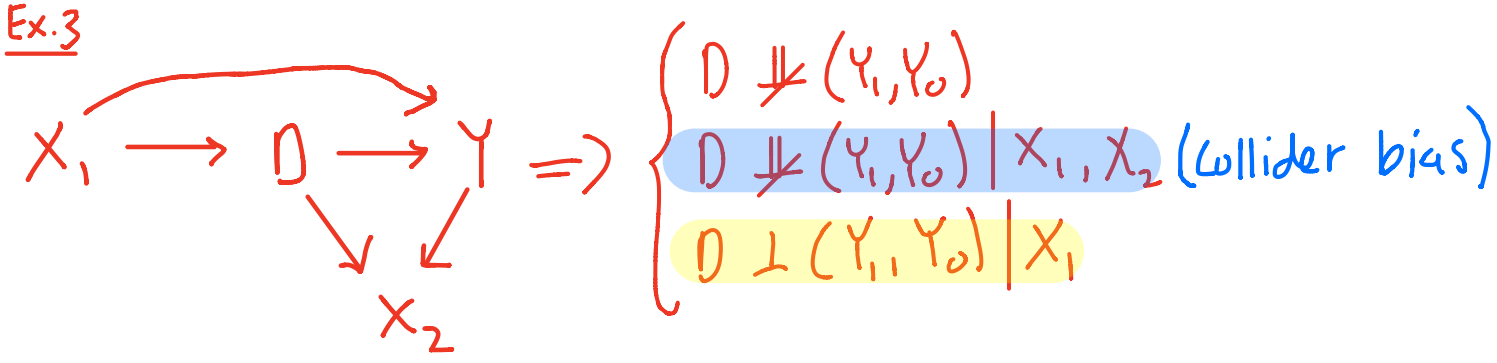
- We will illustrate using examples below on how the DAG can help us in determining the set of conditioning variables such that the treatment will be independent of potential outcomes.



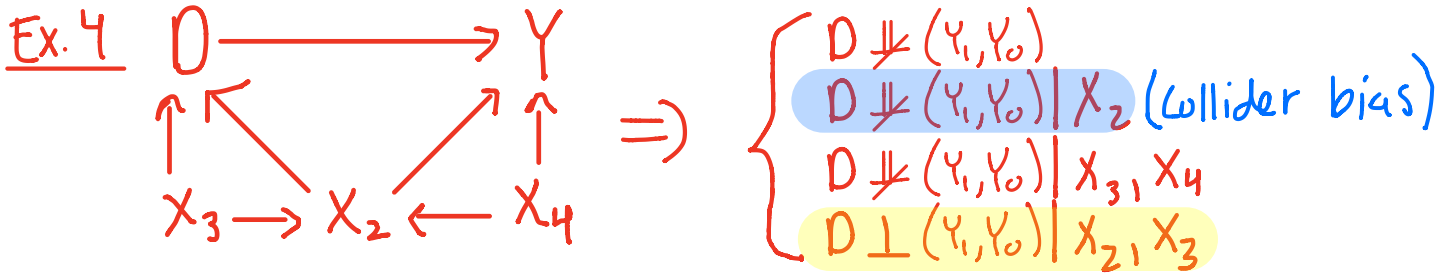
- Without conditioning in the above example there are two backdoor paths 1) D <-- X1 --> Y and 2) D <-- X2 <-- U --> Y. However both of these **open** backdoor paths can be blocked by conditioning on X1 and X2 (none of these controls are **colliders** or **bad controls**).



- There are two **open** backdoor paths in the above example 1) D <-- X3 <-- X1 --> Y and 2) D <-- X3 <-- X2 --> Y. We can block these backdoor paths by conditioning on X1 and X2 or conditioning just on X3.
- Again note that in the above example X1, X2, and X3 are not **colliders** or **bad controls**.



- For the above example there is one backdoor **open** path D <-- X1 --> Y. This is can be **blocked** by controlling for X1.
- There is also a **closed** path D --> X2 <-- Y. If we mistakenly control for X2, since X2 is a **collider** in this path, the path will **open** and hence the direct effect will not be identified.
- Let us consider one more example which is slightly more complex that will also help illustrate **collider bias**.



- There are two **open** backdoor paths: D <-- X2 --> Y and D <-- X3 --> X2 -- > Y. Also there is one initially **closed** backdoor path D <-- X3 --> X2 <-- X4 --> Y.
- Controlling for X2 leads to **collider bias** by **opening** last path. Conditioning on X2,X3 meets CIA.
- To summarize, DAGs can be used to inform researchers the variables they need to control for to satisfy the Conditional Independence Assumption (CIA).

Introducing The "do" Operator

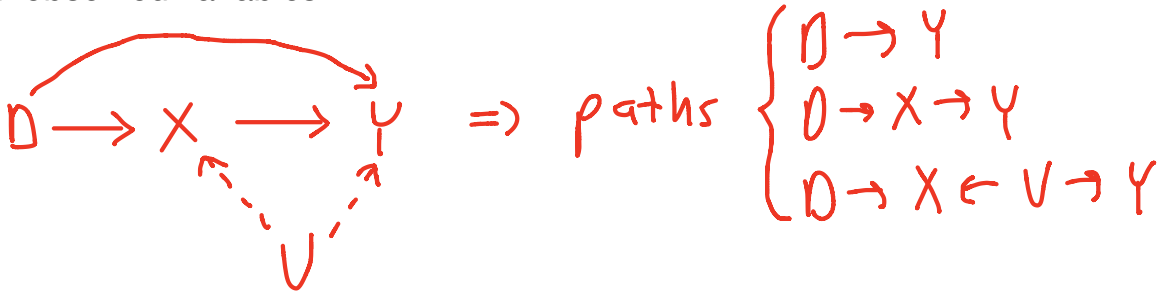
- The do(.) operator marks an action or intervention that has to be completed by the specified units. That is all units perfectly comply to the actions given by the do(.) operator.
- The do(.) operator can be used to formulate the average treatment effect as follows:

$$\hookrightarrow ATE = E(Y_i|do(D_i=1)) - E(Y_i|do(D_i=0))$$

- Note that E(Y|D = 1) - E(Y|D=0) may not equal ATE = E(Y|do(D = 1)) - E(Y|do(D = 0)) in observational data.
- E(Y|do(D = 1)) is analogous to the E(Y1) (average potential outcome if in treatment group), and E(Y|do(D = 0)) is analogous to E(Y0).
- Similarly the marginal distribution of the potential outcome F(Y1) is analogous to F(Y|do(D=1)).
- Identification of the direct effect means that we can learn about E(Y|do(D = 1)) - E(Y|do(D = 0)).

DAG Terminology Review

- Let us summarize our discussion so far by reviewing and also introducing some new DAG terminology.
- A **path** is a sequence of random variables connected with directed arrows.
- **Solid lines** in the DAG indicate that all variables are observed. **Dashed lines** are used for unobserved variables.



- U is not observed as indicated by the dashed lines.
- There are two **backdoor paths** 1) D --> X --> Y and 2) D --> X <-- U --> Y.
- A **collider** is any endogenous variable that is caused by two or more variables.
- X is a **collider** for the path D --> X <-- U --> Y and X is a **non-collider** for path D --> X --> Y.
- U is **parent** of X and Y (arrows pointing from U to both X and Y). Similarly X and Y are **children** of U.
- A **direct** path is connected by the same direction arrows. Other paths are **indirect.**
- There is a **direct** path from D to Y (D --> X --> Y) but no direct path from D to U
- There are two **indirect** paths from D to U: 1) D --> X <-- U and 2) D --> X --> Y <-- U
- Y is a **descendant** of D (**direct** path from D to Y). Similarly D is a **ancestor** of Y.

Conclusion

- DAGs are graphic models that non-parametrically represent causal inference problems.
- Conditioning is not always a good idea for causal identification. Conditioning on colliders will **open** the otherwise **closed** backdoor paths. Conditioning on **bad controls** will explain away portions of the direct effects of interest.
- It is important to emphasize that the relationships in the DAG are determined by models and economic theory. We have to think about where a certain variable is a **collider** or not.
- Collider bias can also arise if we condition on a subset of the population that is a **collider** (sample selection).
- The **backdoor criterion** can be used as a guideline to pick appropriate conditioning variables for causal identification.
- The **do(.)** operator can formulate the ATE and relate DAGs to the potential outcome model.